Neural Networks 167 (2023) 400–414

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations

Amr Farahat [a,b,*], Felix Effenberger [a,c], Martin Vinck [a,b]

[a] Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany
[b] Donders Centre for Neuroscience, Department of Neuroinformatics, Radboud University, Nijmegen, The Netherlands
[c] Frankfurt Institute for Advanced Studies, Frankfurt, Germany

ARTICLE INFO

Article history:
Received 12 December 2022
Received in revised form 7 July 2023
Accepted 13 August 2023
Available online 18 August 2023

Keywords:
Computer vision
Object recognition
Visual cortex
CNNs
Shape representations
Texture bias

ABSTRACT

Convolutional neural networks (CNNs) are one of the most successful computer vision systems to solve object recognition. Furthermore, CNNs have major applications in understanding the nature of visual representations in the human brain. Yet it remains poorly understood how CNNs actually make their decisions, what the nature of their internal representations is, and how their recognition strategies differ from humans. Specifically, there is a major debate about the question of whether CNNs primarily rely on surface regularities of objects, or whether they are capable of exploiting the spatial arrangement of features, similar to humans. Here, we develop a novel feature-scrambling approach to explicitly test whether CNNs use the spatial arrangement of features (i.e. object parts) to classify objects. We combine this approach with a systematic manipulation of effective receptive field sizes of CNNs as well as minimal recognizable configurations (MIRCs) analysis. In contrast to much previous literature, we provide evidence that CNNs are in fact capable of using relatively long-range spatial relationships for object classification. Moreover, the extent to which CNNs use spatial relationships depends heavily on the dataset, e.g. texture vs. sketch. In fact, CNNs even use different strategies for different classes within heterogeneous datasets (ImageNet), suggesting CNNs have a continuous spectrum of classification strategies. Finally, we show that CNNs learn the spatial arrangement of features only up to an intermediate level of granularity, which suggests that intermediate rather than global shape features provide the optimal trade-off between sensitivity and specificity in object classification. These results provide novel insights into the nature of CNN representations and the extent to which they rely on the spatial arrangement of features for object classification.

1. Introduction

The development of Convolutional Neural Networks (CNNs) has led to a revolution in the field of computer vision (Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Yoshua, & Geoffrey, 2015). Machine vision using CNNs has been able to rival human performance in object recognition tasks on large-scale datasets such as ImageNet (He, Zhang, Ren, & Sun, 2016). Moreover, a series of recent works have shown that CNN activations can be used to predict neural activity in the ventral stream of the primate visual system known to be responsible for object recognition (Cadieu et al., 2014; Yamins & DiCarlo, 2016; Yamins et al., 2014). Therefore, there has been a growing interest in developing behavioral benchmarks that evaluate similarities and differences between CNN models and human vision (Geirhos, Narayanappa, Mitzkus, Thieringer, Bethge, Wichmann, & Brendel, 2021; Geirhos, Temme, Rauber, Schütt, Bethge, & Wichmann, 2018; Rajalingham et al., 2018). Crucial to the behavior of these artificial and biological vision systems is their internal representation of objects. The ability of humans to recognize objects based on their abstract shapes (Baker & Kellman, 2018; Biederman & Ju, 1988; Landau, Smith, & Jones, 1988) suggests that the internal representations of objects in the brain must reflect the global structure of objects (Barenholtz & Tarr, 2006; Biederman, 1987). An abstract representation of the global shape of an object requires the encoding of the spatial relations between the set of its local features or parts (Barenholtz & Tarr, 2006; Biederman, 1987). Accordingly, in order to understand the biases that govern the strategies of CNNs performing object recognition, it is central to determine the spatial extent of the diagnostic features CNNs use for object recognition. Moreover, it is equally important to investigate the role that spatial relations play in the construction of these diagnostic features.

Recent studies have shown inconsistent conclusions regarding the reliance of CNNs trained for object recognition on sets of

* Corresponding author.
E-mail address: amr.farahat@esi-frankfurt.de (A. Farahat).

https://doi.org/10.1016/j.neunet.2023.08.021

local features or a global representation of objects (Baker & Elder, 2022; Baker, Lu, Erlikhman, & Kellman, 2018, 2020; Brendel & Bethge, 2019; Geirhos et al., 2019; Jo & Bengio, 2017; Kubilius, Bracci, & Op de Beeck, 2016; Ritter, Barrett, Santoro, & Botvinick, 2017; Tartaglini, Vong, & Lake, 2022). Some studies have shown that CNNs trained for object recognition are biased towards surface statistical regularities (*texture*) (Baker & Elder, 2022; Baker et al., 2018, 2020; Geirhos et al., 2019; Jo & Bengio, 2017). In these studies, CNNs were tested on image datasets that included, for example, low-frequency filtered images (Jo & Bengio, 2017), shape-texture cue conflict stimuli using style transfer (Gatys, Ecker, & Bethge, 2016; Geirhos et al., 2019), deformed silhouettes and other abstract shape images (Baker & Elder, 2022; Baker et al., 2018) and simple geometric shapes (Baker et al., 2020). However, other studies reached different conclusions using other image manipulations or different evaluation methods (Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022). We reckoned that these different conclusions may be due to the hypothesis-driven approach resulting from the choice of the nature of the stimulus datasets and the object classes represented in them. For this reason, we developed a framework for training and testing CNNs that enables us to inspect the shape representations of CNNs by separately controlling the granularity of CNN features (local vs. global) and the spatial relations between them. This approach allows us to take on the question of to what extent the CNN architecture constrains their capacity to learn shape representations and whether CNNs use the spatial relations among features for object recognition.

Previous work has shown that grid-based image scrambling can be used to identify brain areas sensitive to global configurations of objects (Grill-Spector, Kushnir, Hendler, Edelman, Itzchak, & Malach, 1998), expressing characteristic decreases in neural activity with the degree of image scrambling (Grill-Spector et al., 1998; Rainer, Augath, Trinath, & Logothetis, 2002; Vogels, 1999). Image scrambling, however, disrupts not only the spatial relations between object parts but also the shape of the parts themselves (Margalit, Biederman, Tjan, & Shah, 2017). To disentangle these two effects, we developed a feature-scrambling approach that allows us to spatially scramble the pretrained features of CNNs with restricted effective receptive fields (ERFs) (Brendel & Bethge, 2019) without introducing the confounding factors of an image-based scrambling approach. The ERF of a CNN is defined as the set of all pixels that can influence the activity of a unit in its last convolutional layer (Le & Borji, 2017). These features represent diagnostic parts of the objects at the ERF level of granularity. After that, we feed these scrambled features to a follow-up CNN that spatially integrates these features and is trained to recognize the class of objects. Recent work suggests that CNNs with restricted ERF sizes can achieve a performance similar to regular CNNs on ImageNet (Brendel & Bethge, 2019). However, it remains unclear whether these models use the same strategies as regular CNNs to solve the task. Notably, the approximation of regular CNNs performance on ImageNet with CNNs with restricted ERFs implies that CNNs rely on a classification strategy that pools local evidence from separate locations in the image without learning the spatial relations between them. This observation would predict, for instance, that training a follow-up CNN on the pretrained features of a CNN with restricted ERFs should minimally affect performance. It would also predict that spatially scrambling the pretrained input features to the follow-up CNN would not lead to a significant difference in performance to training with the right spatial arrangement of the features. In this work, we tested these predictions on different datasets that comprise texture-rich and texture-less images to examine whether CNNs employ different classification strategies for different datasets. Furthermore, we examined to what extent CNNs with smaller ERFs develop representations similar to CNNs with larger ERFs. Finally, we performed a minimal recognizable configuration (MIRC) analysis (Ullman, Assif, Fetaya, & Harari, 2016) to quantify the minimal image patch sizes required by CNNs to achieve correct classification.

2. Methods

2.1. Datasets

We trained CNNs on three datasets with different feature characteristics: the Sketchy, Animals, and ImageNet datasets. The Sketchy dataset contains 75,471 human-drawn sketches spanning 125 classes (Sangkloy, Burnell, Ham, & Hays, 2016). Each sketch is a textureless, black-and-white bitmap graphic that only contains information about the contours of objects without any surface proprieties, and sketches have a high degree of intra-class variability (Fig. 1c). The Animals dataset consists of 37,322 color images spanning 50 classes (Xian, Lampert, Schiele, & Akata, 2019) (Fig. 1b). The well-known ImageNet dataset contains 1.2M color images across 1000 classes (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009) that span different animals and man-made artifacts.

2.2. Models

We created residual CNNs (He et al., 2016) with ERFs of variable sizes (Table 1) by changing the size of the filters of different residual units across layers (Brendel & Bethge, 2019). The residual CNNs consist of 4 blocks that contain 2, 3, 3, and 2 residual units, respectively. Each residual unit consists of 3 convolutional layers: The first and last layers always have filters of size 1×1 and the filter size of the middle layer varies according to Table 1. Adjusting the filter size of the residual units results in models with ERFs of either 11, 23, 47, 95, or 227 pixels squared in the last layer. We refer to these models by their ERF sizes, writing ERF23 for a network with an ERF of size 23×23 pixels. Note that since our input images are always of size 224×224 pixels, only the model ERF227 has units in the last convolutional layer with ERFs covering the entire image, before features are globally averaged across spatial locations in the penultimate layer.

2.3. Feature-scrambling approach

For the feature-scrambling approach, we build CNN models that are composed of two sub-networks, a *base network* and a *follow-up network* (Fig. 1d). The base network transforms the image to high-level feature maps of a given size by being trained on image classification in a standalone way. These pretrained features are then fed into a follow-up network. The follow-up network then further transforms these feature maps in a series of convolutional layers. Finally, features are pooled in a location-discarding way in a global average pooling layer and then a Softmax classification layer. This approach allows us to independently examine the granularity of features used by CNNs for object recognition and to determine to what extent the spatial relations among them contribute to their performance.

We used networks with different ERFs as base networks. We trained them separately for image classification and then detached the fully connected classification layer and the global average pooling layer of the trained network and used it with frozen weights as the base network in our feature-scrambling approach. Subsequently, we attached the follow-up network such that it receives the features of the pretrained base networks as inputs in either a scrambled or unscrambled way. Specifically, for the unscrambled case, we passed the feature maps unchanged to

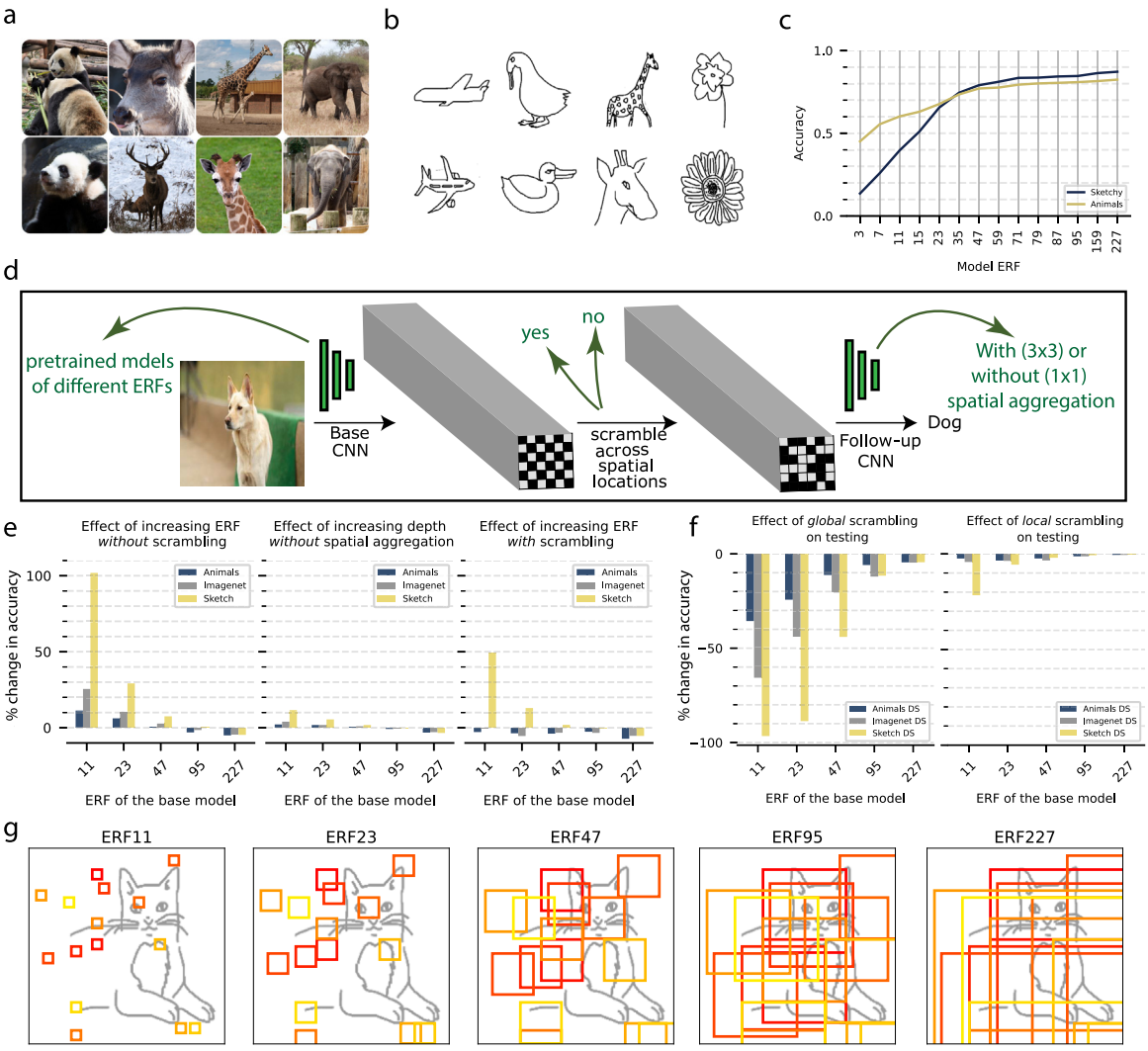

Fig. 1. Feature scrambling during training and testing. (a, b) Example images for the Animals and Sketchy datasets, respectively. (c) CNN performance as a function of the ERF, separately for the Sketchy and Animals datasets. (d) A schematic for the feature-scrambling approach. (e) Effects of adding the follow-up network to the pretrained base networks either with spatial aggregation without scrambling (left), with spatial aggregation with scrambling (right), or without spatial aggregation (middle). (f) Effect of global and local feature-scrambling on the testing performance of the base + follow-up models with spatial aggregation without scrambling. (g) A schematic depicting the ERF of random artificial neurons in the last convolutional layer of models of different ERFs.

Table 1
Architecture details for our ResNets of different ERFs.

Blocks	Residual units	Feature maps	Stride	Filter sizes				
				ERF11	ERF23	ERF47	ERF95	ERF227
Block 1	2	128	2	3,3	3,5	3,5	3,5	5,5
Block 2	3	256	2	1,1,1	3,1,1	3,3,5	3,3,5	5,5,5
Block 3	3	512	2	1,1,1	1,1,1	1,1,1	3,3,3	5,5,5
Block 4	2	1024	1	1,1	1,1	1,1	1,1	5,5

the follow-up network. For the scrambled case, we generated random indices once and used them to permute the feature vectors across spatial locations. The follow-up network is a residual block formed of four residual units. We differentiated between two types of follow-up networks: *with* or *without* spatial aggregation: (1) A follow-up network *with* spatial aggregation has a stride of 2 for its first two residual units and filter size 3×3 for all its residual units. (2) A follow-up network *without* spatial aggregation is formed exclusively of convolutional layers with filter size 1×1 and no down-sampling. In summary, for each of our base networks (ERF11, ERF23, ERF47, ERF95, and ERF227), we trained 3 models depending upon: (1) the type of the follow-up network (with or without spatial aggregation); (2) scrambling the features between the two sub-networks or not.

The considered models can be summarized as follows:

- **Base:** only the base network trained in a standalone way.
- **Base + Follow-up without scrambling:** the model is formed of the pretrained base network plus the follow-up network with spatial aggregation and without feature-scrambling.
- **Base + 1×1 Follow-up without scrambling:** the model is formed of the pretrained base network plus the follow-up

network without spatial aggregation and without feature-scrambling. This model serves as a control for the significance of increasing the ERF of the model by adding the follow-up network.

- **Base + Follow-up with scrambling:** the model is formed of the pretrained base network and the follow-up network with spatial aggregation and with global feature-scrambling during training.

Additionally, the **Base + Follow-up without scrambling** models were tested while the input features to the follow-up network were randomly scrambled either globally or locally.

2.3.1. Training

All simulations were performed using the TensorFlow library (Abadi et al., 2015). We used stochastic gradient descent with momentum = 0.9 to update the weights with initial learning rate = 0.01 for the first 10 epochs followed by exponential decay for the rest of training. For the ImageNet dataset, we trained for 50 epochs, and for the animals and Sketchy datasets, we trained for 75 epochs.

During training, for non-square images, we first cropped the central square portion of the image with the shortest dimension of the image to keep the aspect ratio of the objects constant before resizing the images to 256 × 256 pixels. We then applied minimal data augmentation in the form of random right and left horizontal flipping of the images, followed by random cropping of 224 × 224-pixel patches used for training. During testing, after centrally cropping the images, we resized them to 256 × 256 pixels, and then we cropped the central 224 × 224-pixel patch.

2.4. Representational similarity analysis (RSA)

We used RSA to investigate the representations of the CNNs of different ERFs (Nili et al., 2014). To avoid the results being biased to the number of classes in each dataset, we sampled 50 random classes from each dataset (the lowest number of classes in the three datasets). Then we sampled 8 random images from each class for a total of 400 images, ran them through all the models of different ERFs, and extracted the activations of the last convolutional layer of each residual unit ($n = 10$), the global average pooling layer (GAP) and the Softmax layer. For the Sketchy and Animals datasets, we averaged the layers' RDMs across 5 repetitions of random initialization. We created the representation dissimilarity matrix (RDM) for each layer by computing the pairwise correlation distance for its activations (400 × 400 matrix). Next, we computed a second-order RDM for all the layers of the models (60 × 60 matrix) by computing the correlation distance between the upper triangle of the layers' RDMs. For visualization purposes, we used multi-dimensional scaling (MDS) to reduce the dimensionality of the second-order RDM to two dimensions.

2.5. Minimal recognizable configurations analysis (MIRC)

We adopted the MIRC analysis (Ullman et al., 2016) previously used for humans for CNNs. MIRC analysis is a recursive process that search for the smallest image patches that still yield a correct classification result. MIRC analysis starts with a given, correctly classified image of class c. Starting from the whole image as one patch, four descendant patches are created from each patch. Each descendant batch spans 75% of the height and width of the patch at the previous level starting from one of the four corners (Fig. 4a). Each patch is then upsampled using bilinear interpolation to 224 × 224 pixels to match the input size of the models. The recursive subdivision process continues for each patch as long as the patch is still correctly classified as belonging to class c. Subdivision stops once the classification of a patch is no longer correct. This process defines a tree structure and the leaves of the tree are the MIRCs. The level of a leaf node in the tree is referred to as the level of the MIRC it represents. By construction, the higher the MIRC level, the smaller the patch of the image used for classification.

3. Results

3.1. Feature scrambling during training and testing

We trained CNNs of different ERF sizes on three different datasets: the Sketchy (Sangkloy et al., 2016), the Animals (Xian et al., 2019) (Fig. 1a–b), and the ImageNet (Deng et al., 2009) dataset. Example ERFs of five models are shown in Fig. 1g. We note that the ERF is a theoretical upper limit on the set of pixels that can activate a given unit, and that not all pixels of the ERF necessarily activate the corresponding deep unit, depending on connection weights. We found that CNN performance increased with ERF size for both the Sketchy and Animal datasets, with a visible saturation for larger ERFs. However, CNN performance depended more strongly on the ERF size for the Sketchy dataset than for the Animals dataset (Fig. 1c). Because changing the filter sizes across models will also induce changes in the number of trainable parameters in the models and consequentially their expressive capacity, we performed a control experiment in which we created wide networks with small ERFs but matched the number of parameters of the network with the largest ERF (ERF227). We found a slight increase in accuracy, but the models still showed a substantial reduction in performance compared to the corresponding network with a large ERF (Fig. A.1).

The dependence of the classification performance on ERF size suggests that the network's ERF has a major impact on object recognition, especially for textureless datasets such as the Sketchy dataset. One explanation for the observed performance increase could be that CNNs with large ERFs can learn to exploit relatively large-scale features, which are especially important for texture-less datasets. However, the comparison between networks with large ERFs and small ERFs does not yet provide direct evidence that CNNs with large ERFs rely on large-scale shape features. For example, it is possible that the pooling in large ERFs does not take into account the spatial configuration among the features. Instead, the network might just accumulate local evidence in a different manner than networks with smaller ERFs. This reasoning suggests that in order to investigate the network's sensitivity to the spatial configuration of features, it is necessary to distort (i.e. scramble) the spatial arrangement of features and then test the impact of this distortion. Importantly, this scrambling should be done at the level of the network features rather than at the image level, as the latter often leads to confounding high-contrast image artifacts. Specifically, we took the following approach:

(1) We trained a network with a small ERF size on an object recognition task. We call this the base CNN, which was not further modified.

(2) We then trained a follow-up network, which received input from the last convolutional layer of the pretrained base CNN. These pretrained input features represent diagnostic features of certain granularity depending on the ERF of the base CNN i.e. object parts at different scales. The follow-up network has an ERF that covers the entire image. We observed that adding the follow-up network led to an increase in performance compared to the base network. Consistent with the ERF survey experiment (Fig. 1c), the increase in performance was relatively small for the ImageNet and Animals datasets but was large for the Sketchy

dataset for base networks with smaller ERFs (Fig. 1e, left panel). Absolute performances are shown in Fig. A.2.

(3) To rule out the possibility that the observed performance increase for such stacked networks was just caused by increasing the depth of the model by appending the follow-up network, we trained a follow-up network that consisted only of 1×1 convolutions without strides to prevent spatial aggregation. We observed only a slight increase in accuracy for all datasets (Fig. 1e middle), which shows that spatial aggregation of inputs was crucial for the observed performance boost (Fig. 1e left).

(4) To examine whether the spatial configuration of features mattered, we trained the same follow-up networks after spatially scrambling the features in the last convolutional layer of the base network. We used a fixed spatial permutation (i.e. scrambling) of these features that was constant during training. We observed a smaller increase in performance for the Sketchy dataset for ERFs 11 and 23 (Fig. 1e right). Furthermore, no further increase in accuracy could be observed for the Animals and ImageNet datasets in this case (Fig. 1e right). Taken together, these findings suggest that CNNs can learn to utilize the configuration of spatially distant features when constructing more complex features in subsequent layers, especially for datasets in which shape is expected to be critical for object classification.

(5) As a complementary approach to the fixed scrambling during training, we also performed random feature scrambling during testing. As before, the scrambling was again done at the last convolutional layer of the base network. As predicted, we observed a general decrease in the accuracy of the models with spatial aggregation (base + follow-up) when the features were globally scrambled during testing (Fig. 1f left). This effect depended strongly on the dataset, with a relatively weak effect for the Animals dataset and a very strong effect for the Sketchy dataset. Moreover, the performance reduction was particularly pronounced for models with small ERFs that exclusively encode local features of fine granularity before the scrambling is done. It is worth noting that this effect cannot be simply explained by the type of the dataset (sketches versus natural images) since the reduction in performance varied substantially between the Animals and ImageNet datasets, even though both consist of natural images.

(6) As a control, we also performed a "local" scrambling, in which the features were scrambled only at neighboring locations. The reduction in performance with local scrambling was much weaker compared to global scrambling, indicating that the loss of performance with global scrambling is due to the distortion of the global configuration of the features, not the confounding effects of the scrambling process itself.

Together, these results highlight the importance of the granularity of features and their spatial configurations for object recognition, especially for datasets in which texture is less informative. In other words, models with larger ERF can extract more coarse-grained features, which are more diagnostic for the object class, i.e. have higher accuracy and are less susceptible to scrambling. These coarse-grained features are diagnostic on their own and do not need to be spatially integrated to construct more complex features in subsequent layers (the follow-up network). However, the granularity of these features differs between datasets.

3.2. Variability of classification strategies between classes in ImageNet

Depending on the dataset, we observed different effects of ERF sizes and feature scrambling on network classification performance. Changing the ERF size had the weakest effect on performance for the Animals dataset and the strongest effect for the sketches dataset, with ImageNet in between (Fig. 1e left). The strongest effect of feature scrambling was observed on the Sketchy dataset, followed by ImageNet and then the Animals dataset, which was least affected by feature scrambling (Fig. 1f left). These findings can be explained by the image statistics in the different datasets. Two extremes are given by the Animals and Sketchy dataset: While images in the Animals dataset can already be classified using local textural features, pictures in the Sketchy dataset require the integration of spatially distant features for classifications. For ImageNet, the classification may allow for different class-specific strategies (e.g., animals vs. man-made artifacts). To test the hypothesis that CNNs use different classification strategies for different ImageNet classes, we used the feature-scrambling approach described above. As a measure of how CNN classification performance is affected by global feature scrambling, we consider the scrambling ratio as the ratio of class $f1$ scores before and after scrambling. A high scrambling ratio indicates that a class is not sensitive to feature scrambling (which we call scrambling-insensitive), and a low value indicates sensitivity to scrambling (which we call scrambling-sensitive). This ranks the classes according to their sensitivity to the global spatial feature configuration in the last CNN layer of the base network (Fig. 2a-c). For this analysis, we only considered classes that the model reliably classified before scrambling ($f1 > 0.75$).

As hypothesized, the least scrambling-sensitive classes predominantly express characteristic surface patterns (texture) such as the rapeseed, brain coral, and zebra classes (Fig. 2a and b in blue for base models ERF11 and ERF23 respectively). Scrambling-sensitive classes, on the other hand, were not found to express such characteristics textures, such as the water tower, electric locomotive, and horse cart classes (Fig. 2a and b in yellow for base models ERF11 and ERF23 respectively). We hypothesized that the variability in scrambling sensitivity was due to the intrinsic properties of the classes and their performance at low ERFs, rather than to the scrambling operation itself. In fact, we found that classes with high scrambling sensitivity only exhibited this high sensitivity for models with small ERFs (Fig. 2c right). However, the scrambling sensitivity of classes was found to be mostly independent of ERF size (Fig. 2c left). To confirm that this effect is a consequence of the heterogeneity of the ImageNet dataset and not the ordering process, we repeated the same analysis for the Animals dataset and did not observe such substantial variability in the scrambling ratios among classes, e.g., for the base model ERF11, scrambling ratios ranged from 0.05 to 0.91 and from 0.61 to 0.96 for ImageNet and Animals datasets respectively. We furthermore found that the set of the least scrambling-sensitive classes is mostly consistent across models (Fig. 2d left). This is in contrast to the set of the most scrambling-sensitive classes (Fig. 2d right). Thus, the performance of scrambling-sensitive classes depends more on the models' ERFs and, therefore, relies on features of coarser granularity.

Therefore, we hypothesized that the scrambling ratio should predict the performance increase from the ERF11 to the ERF227 network (Fig. A.2), as well as the performance increase obtained by adding the follow-up network to the pretrained base network (Figs A.2 and Fig. 1e). Indeed, the performance increase for ERF227 compared to ERF11 and ERF23 was greater for scrambling -sensitive than for scrambling-insensitive classes (Fig. 2e). Specifically, the performance ($f1$ score) of the model ERF227 on the 20 most scrambling-sensitive classes was higher than that of all other models (ERF11, ERF23, ERF47, and ERF95) in a statistically significant way according to the Wilcoxon signed-rank test. In contrast, for the 20 least scrambling-sensitive classes, the performance of the ERF227 model was only significantly higher than the models ERF11, ERF23, and ERF47, but not ERF95. Similarly, the performance increase caused by the addition of a follow-up network was larger for scrambling-sensitive classes than for

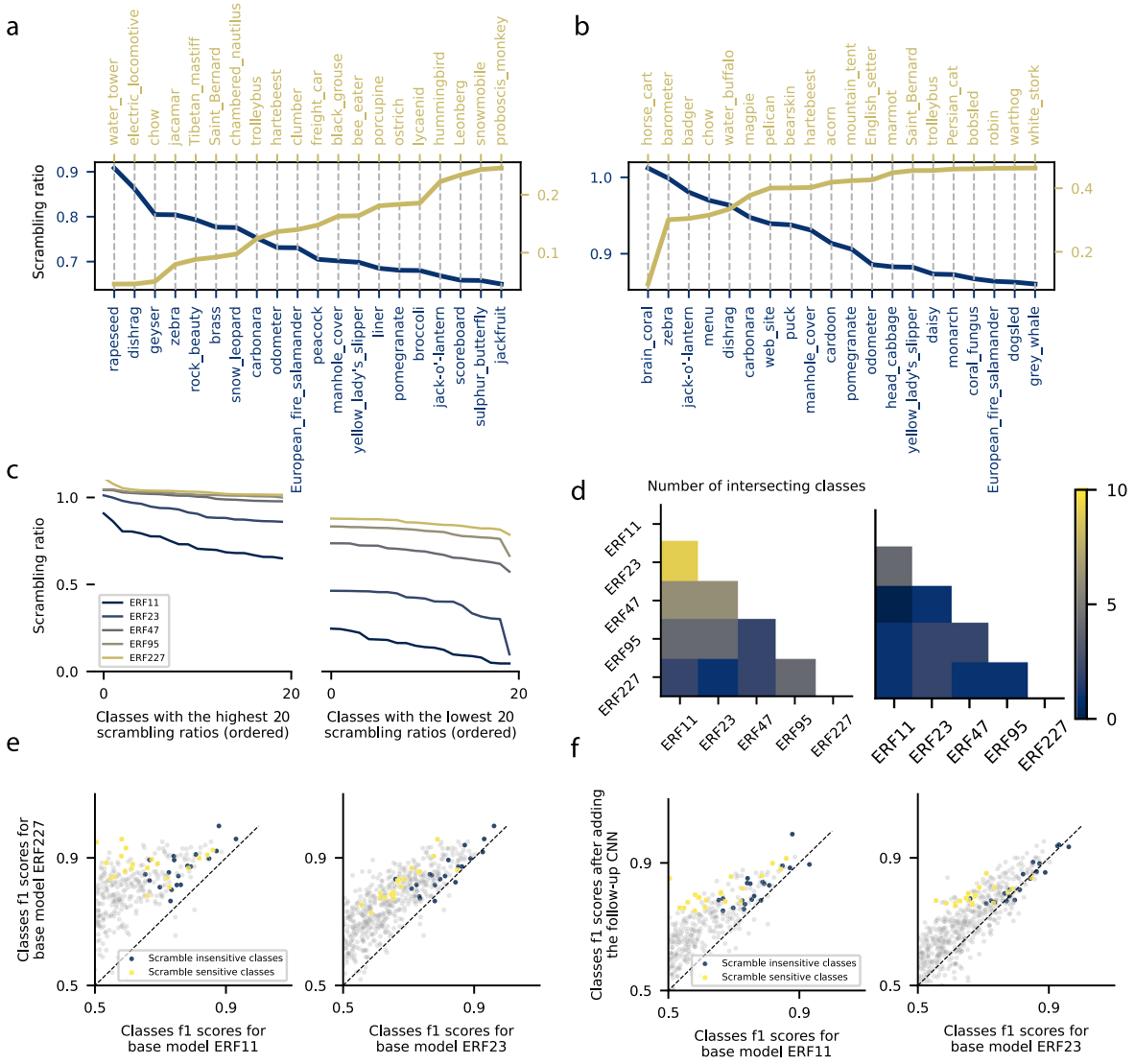

Fig. 2. (a, b) The 20 least (in blue) and most (in yellow) scrambling-sensitive ImageNet classes for the models ERF11 (a) and ERF23 (b). (c) Scrambling ratios of the 20 least (left) and most (right) scrambling-sensitive ImageNet classes for models of different ERFs. High and low values of the scrambling ratio indicate that feature-scrambling has minor and major effects on class performance, respectively. (d) Number of the intersecting classes for the 20 least (left) and most (right) scrambling-sensitive ImageNet classes among models of different ERFs. (e) $f1$ performance scores of ImageNet classes for ERF11 and ERF23 models against ERF227 model. In blue and yellow are respectively the least and most scrambling-sensitive classes. (f) $f1$ performance scores of ImageNet classes for the base model vs. base model after adding the follow-up network. In blue and yellow are respectively the least and most scrambling-sensitive classes.

scrambling-insensitive classes (Fig. 2f). In particular, increasing the ERF of the models by adding the follow-up network led to a statistically significant increase in the performance of the 20 most scrambling-sensitive classes for the models ERF11, ERF23, ERF47, and ERF95. For the 20 least scrambling-sensitive classes, it only led to a statistically significant increase in performance for the models ERF11 and ERF23.

3.3. Representation similarity analysis

Next, we investigated the role of ERF size on the classification strategies used by CNNs. We used representation similarity analysis (RSA) (Nili et al., 2014) to test whether CNNs of different ERF sizes develop comparable representations, reflecting similar or different classification strategies (Fig. 3a–c; see Section 2). For each layer, we computed a representation dissimilarity matrix (RDM) by computing the pairwise correlation distance on the activations resulting from different images. We then computed the dissimilarity (using the pairwise correlation distance) of the RDMs between all layers of all models, thus indicating the similarity of the representations between different layers of different models. To facilitate visualization, we employed multi-dimensional scaling to reduce the dimensionality of the RDM so that each point in the 2-d space represents a layer in a model and connected the layers of each model with a solid line of a different color (Fig. 3a). We observed that models with comparable ERFs are closer in the low dimensional space (Fig. 3a), indicating that the distances among the corresponding layers of the models depend on the models' ERF.

To further investigate whether CNNs with small ERFs use classification strategies similar to those of standard CNNs with large ERFs, we correlated the RDMs for all models with the RDM for the ERF 227 model. Specifically, we computed the variance explained (R^2) between the RDMs of the ERF227 model and the

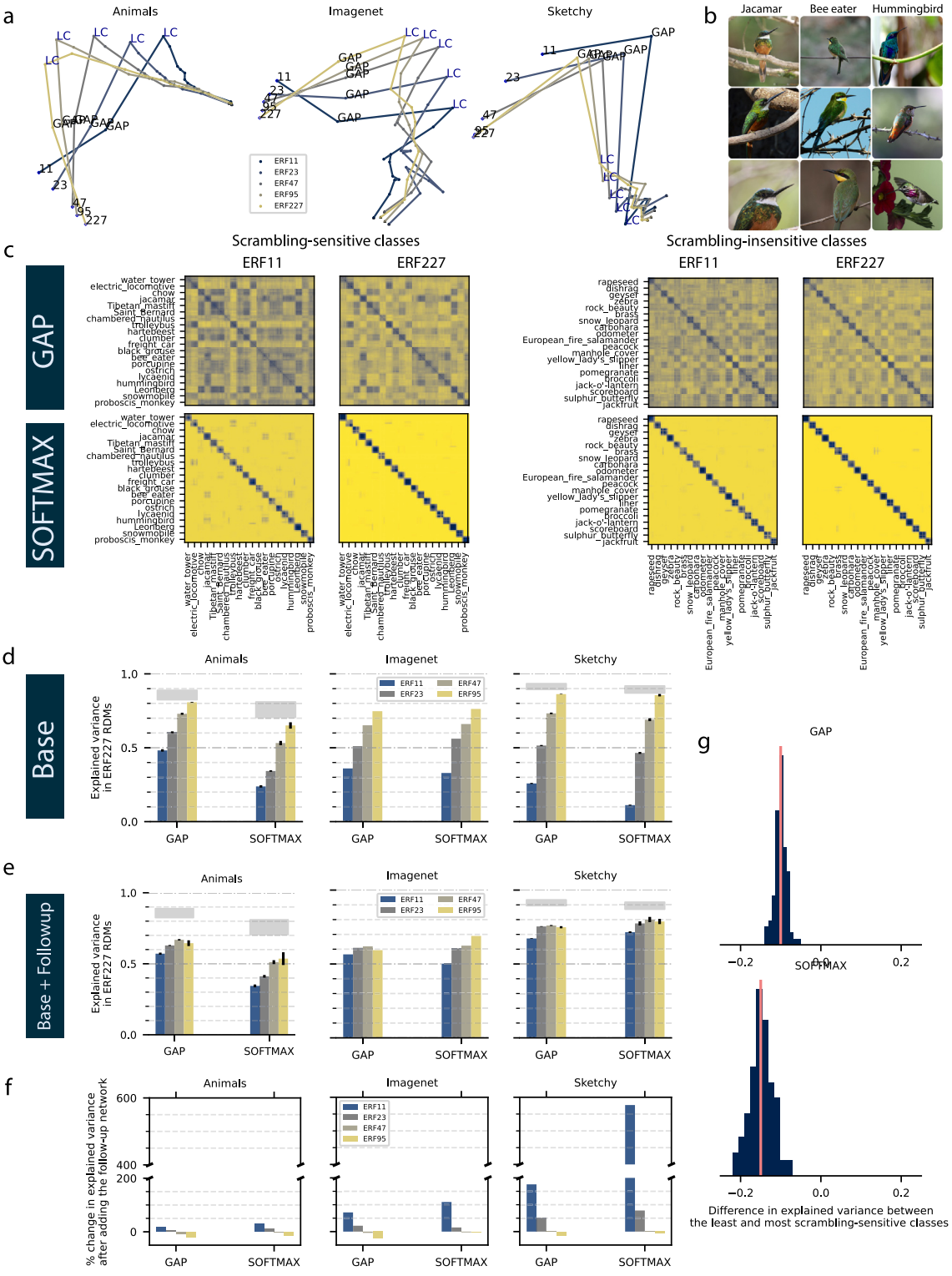

Fig. 3. (a) Representation trajectories for five CNNs with different ERFs trained on 3 different datasets. For the Sketchy and Animals datasets, we averaged the layers' RDMs (Representational Dissimilarity Matrices) of 5 training iterations of each model of a certain ERF size before computing the second-order RDM of all layers. LC: last convolutional layer. GAP: Global Average Pooling layer. ERF number indicates the classification layer of the corresponding model. (b) Each column shows three examples from the ImageNet dataset for three bird classes. (c) RDMs of the global average pooling (GAP) and Softmax layers for the models ERF11 and ERF227 computed separately on the 20 least and most feature-scrambling sensitive ImageNet classes as estimated using the ERF11 model and the feature-scrambling approach. We sampled 20 images randomly from each class so each RDM is 400 × 400 (better viewed digitally). (d) The amount of explained variance (R^2) by the GAP and Softmax layers' RDMs of models with different restricted ERFs in the RDMs of the ERF227 model. (e) The amount of explained variance (R^2) by the GAP and Softmax layers' RDMs of models with different restricted ERFs after adding the follow-up network in the RDMs of the ERF227 model. (f) Percentage change in the amount of explained variance by RDMs of models with different restricted ERFs in the RDMs of the ERF227 model after adding the follow-up network that increases the ERF of the models to cover the whole image. (g) The distributions of the difference in explained variance by ERF11 model RDMs in ERF227 model RDMs between scrambling-sensitive and scrambling-insensitive classes. RDMs were computed by randomly sampling images separately from the scrambling-sensitive and scrambling-insensitive classes. The number of repetitions is 100.

RDMs of the models with smaller ERFs (Fig. 3d). This was done separately for the Global Average Pooling (GAP) and Softmax layers for the three datasets. For both GAP and Softmax, we observed a gradual increase in the amount of explained variance with ERF size, i.e., models with small ERFs are more dissimilar to the ERF227 model. Moreover, the amount of explained variance depended on the dataset: The Sketchy dataset had the lowest amount of explained variance for models with small ERF, followed by ImageNet and the Animals datasets. This result agrees with the differences between datasets in terms of the models' classification performance (Fig. 1e). We repeated the same analysis after adding the follow-up networks to the base models, which in each case increased the ERF to cover the whole image (e.g. 235 pixels^2 for ERF11 base model) (Fig. 3e). We noticed an increase in the amount of explained variance after adding the follow-up network, especially on the Sketchy dataset and for the models with small ERFs (Fig. 3f). Again, there was only a minor and intermediate increase for the animals and ImageNet databases, respectively. This supports the notion that CNNs can deploy different classification strategies depending on their ERF.

Furthermore, according to our feature-scrambling analysis, CNN classification strategies should also differ among object classes even within the same model. Therefore, we hypothesized that the explained variance between ERF11 and ERF227 should differ between the scrambling-sensitive and scrambling-insensitive classes. In particular, we expected that the explained variance should be smaller for scrambling-sensitive classes because, for those classes, one expects more spatial integration. For that purpose, we selected the 20 most and least scrambling-sensitive classes of the ImageNet dataset as determined by our feature-scrambling approach for the base model ERF11 (Fig. 2a), randomly selected 20 images from each class, passed them through the models ERF11 and ERF227, computed the RDMs of the GAP and Softmax layers for each model (ERF11, ERF227) and condition (scrambling-sensitive, scrambling-insensitive) separately (Fig. 3c). We repeated the process 100 times and each time we calculated the variance explained in model ERF227 RDMs by model ERF11 RDMs for both conditions. We subtracted the variance explained in the condition of the scrambling-insensitive classes from the variance explained in the condition of the scrambling-sensitive classes to create a distribution of the difference in the variance explained by model ERF11 in model ERF227 between the scrambling-sensitive and scrambling-insensitive classes (Fig. 3g). Indeed, we observed the expected difference (Fig. 3g). Additionally, by visual inspection, the difference between the RDMs of the models ERF11 and ERF227 calculated on the scrambling-sensitive classes is especially pronounced in the off-diagonal part of the matrix, which represents the similarity among the inter-class pairs of images (Fig. 3c left two columns). We hypothesized that the reason behind this difference is that the ERF11 model extracts lower-level features that are not indicative of a specific class, but rather shared among multiple classes. For example, we observe these blocks of low dissimilarity in Fig. 3c (most lower left panel) between the class jacamar and the classes bee-eater and hummingbird, which have shared color and local features (Example samples of each of the three classes are shown in Fig. 3b). Together, these results further support our conclusion that the granularity of features used by CNNs (which in terms are determined by their ERF sizes) plays a crucial role in their ability to perform object recognition. Moreover, the granularity of the CNN features is determined not only by its ERF but also by the statistics of the images in the datasets, separately for each class. Although more coarse-grained features can be more reliable for object recognition, they are only exploited by CNNs when needed e.g. the Sketchy dataset and scrambling-sensitive classes in ImageNet. This agrees with the simplicity bias in CNNs (and more generally all neural networks) when trained with a gradient-based learning rule: Networks tend to become selective to the easiest (and most local) features that allows them to solve the classification task at hand.

3.4. Minimal recognizable configurations (MIRCs) analysis

The results so far suggest that CNNs recognize objects based on features that vary in their granularity depending on the dataset and the object class. For datasets and object classes that have relatively little or no texture information, CNNs can learn to construct diagnostic features of coarser granularity from more fine-grained features by exploiting the spatial relations between them. This raises the following questions: (1) What is the spatial extent of these coarse features and spatial relations learned by CNNs? (2) What is the advantage of more coarse-grained features over more fine-grained features for object recognition? The feature-scrambling results shown above indicate that even for the Sketchy dataset, increasing the ERF of the base models beyond 47×47 had a limited effect on performance. This result suggests that the features required for reliably recognizing objects are still predominantly local, i.e., they span maximally about 4%–5% of the image.

To further test the reliability of the features utilized by models of different ERFs and visualize them in the image space, we performed a MIRC analysis. MIRC analysis tests the ability of the models to categorize images based on localized image patches by searching for the minimal (i.e. smallest) feature configurations in the image that are still correctly recognizable by the models. We searched for the MIRCs of each image in the test dataset of the Sketchy and Animals datasets, and randomly sampled one-third of the images in the test dataset of the ImageNet dataset. For each image, we cropped 75% of the image starting from each corner so that each image yields 4 descendants (Fig. 4a). We then upsampled each descendent crop to the original image size (224×224) and used the model to predict its object class. We repeated the process for each descendant that was correctly classified by the model until we reached the image that was correctly identified by the model but had no correctly classified descendants. This image was declared a MIRC and its level in the search tree defines its size, i.e. the deeper (higher) the level, the smaller the image patch.

In Fig. 4b, we show examples of MIRCs generated from three different images for the zebra class from the three datasets and their deepest MIRCs that have the highest classification probabilities using the ERF227 and ERF11 models. These examples show that on the one hand, the ERF227 model was able to classify the image with high classification probability by relying exclusively on relatively local features, i.e. the zebra's face or stripes. On the other hand, the ERF11 model required larger image patches for successful classification, especially on the Sketchy dataset. This seems to indicate that the model with the larger ERF actually requires a much smaller part of the image to reach the correct classification as compared to the model with the smaller ERF.

To verify whether this finding holds in general, we computed the histograms of the deepest MIRC levels for each image for all datasets and models (Fig. 4c–e). We observed for the base models a dependence between ERF size and maximal MIRC levels, i.e., the larger the ERF size of the CNN, the higher its maximal MIRC levels (i.e. a smaller part of the image was sufficient to classify) (Fig. 4c). By contrast, networks with smaller ERFs typically cover a larger part of the image or the entire image for classification. We found this dependence to be dataset-specific. The difference between ERF227 and ERF11 was largest for the Sketchy dataset and smallest for the Animals dataset. The difference between ERF227 and models with smaller ERFs was reduced after adding

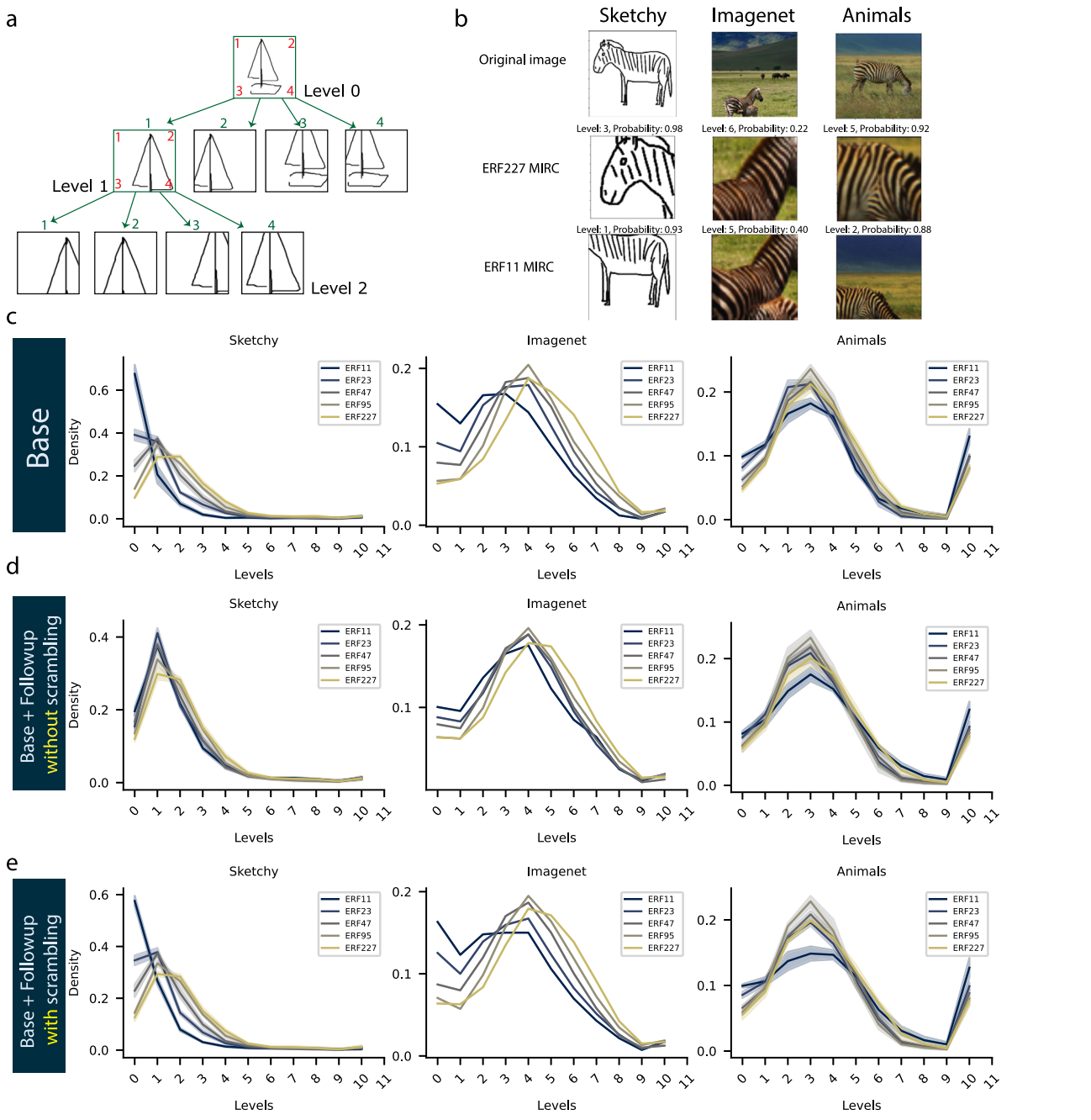

Fig. 4. (a) Illustration of the MIRC procedure. Each image patch yields four descendants. Each descendant is a 75% crop starting from one of the four image patch corners. Each green numbered descendant patch corresponds to the red equivalently numbered corner of the parent patch (See Section 2.5). (b) Example MIRCs for three different images (first row) of the zebra class from three datasets (Sketchy, ImageNet, and Animals) for the models ERF227 (second row) and model ERF11 (third row). The MIRCs shown are the MIRCs with the highest probability (confidence) among the MIRCs of the highest level of that image. (c, d, e) Distribution of the maximum MIRC level for each correctly classified image in the test dataset for the Sketchy, ImageNet, and Animals datasets, respectively for the base networks of different ERF sizes (c), after adding the follow-up network without scrambling (d), and after adding the follow-up network with the spatial scrambling of its input features (e). For the Sketchy and Animals datasets, the histograms are averaged over 5 training iterations. The shaded area represents the standard deviation. The high frequency of images with MIRCs of level 10 in the Animals dataset is because of the images that belong to the classes that the models usually predict when the correct class cannot be identified.

the follow-up network without spatially scrambling the features (Fig. 4d). However, the difference was not affected when a follow-up network was added after spatially scrambling the features during training (Fig. 4e). The effect of feature-scrambling on the distribution of the levels of MIRCs demonstrates the different strategies CNNs can employ for object recognition. On the one hand, spatial integration of features without scrambling led the follow-up networks to be able to construct and be selective to more reliable coarse-grained features than the base models. Subsequently, these models (base + follow-up) had smaller MIRCs than their base models. On the other hand, spatially scrambling the features before feeding them to the follow-up networks prevented them from exploiting the spatial relations between the features to construct more reliable coarse-grained features. Therefore, the follow-up networks were only able to learn the set of more fine-grained features that correlates with the target class. Subsequently, these models retained the relatively large-sized MIRCs of their base models.

To visualize the features required for recognizing a certain class, we obtained latent representations for all MIRCs of all images of a given class using the model. We then used the k-means algorithm to group the latent representations into 5 clusters. In Fig. A.3, we show examples for the horse and eyeglasses classes of the Sketchy dataset for the model ERF227. For each cluster, we show the eight MIRCs that are the closest to the cluster center and originate from distinct images. We observe that each cluster is composed of MIRCs that represent visually similar features. For example, we observe clusters representing hair, the side view of the head, and leg features for the horse class Fig. A.3. For the eyeglasses class, we can identify a cluster containing double-lined frames, one for thin frames, and one for reflective glass features.

4. Discussion

Despite the exceptional performance of CNNs in object recognition tasks (He et al., 2016; Krizhevsky et al., 2012), the nature of their representations is still poorly understood. One aspect of the learned object representations in CNNs is whether they are capable of encoding the global shape of objects. Global shape representations describe objects in the form of both their diagnostic features and the spatial arrangement of these features (Barenholtz & Tarr, 2006; Biederman, 1987) in contrast to models in which the presence of these features can serve alone as evidence for object identity without encoding the spatial relations between them (Edelman, 1993; Wallis & Rolls, 1997). There exists a wide range of visual features e.g. contours, textures, colors, or object parts. We used features of pretrained CNNs of restricted ERFs to represent the diagnostic local features (Brendel & Bethge, 2019). By comparing the two conditions of training a follow-up network on top of these local features either *with* or *without* scrambling of the spatial locations of the features, we could assess the amount of additional information that CNNs can extract by exploiting the spatial relations between features. Moreover, by examining the MIRCs of CNNs, we were able to evaluate the spatial extent of spatial relations learned by CNNs for object recognition.

It has recently been reported that CNN representations may be mostly local (Baker et al., 2020; Brendel & Bethge, 2019) and consequently more biased toward object surface regularities (Geirhos et al., 2019; Jo & Bengio, 2017) than the global form of objects. This led to the hypothesis that they might not be capable of representing spatial relationships among features (Baker & Elder, 2022; Baker et al., 2018). In contrast to conclusions drawn in other works, our analysis allows us to provide the following more nuanced view: (1) We provide evidence that CNNs are capable of using relatively long-range spatial relationships for object classification, especially for textureless datasets (such as sketches). This finding is supported by several analyses, including a new scrambling approach in which we perturbed spatial relations between features within the CNN, and a systematic investigation of how CNN performance is impacted by different effective receptive field sizes. (2) We show that CNNs use different strategies for different datasets, rather than one unified strategy (e.g. pooling evidence based on local texture). Notably, we found that classification strategies can vary even between classes within the same dataset. These strategies differ in the granularity of the features used and in the degree of reliance on the spatial relations between them. This suggests that there is a continuous spectrum of CNN strategies, ranging from exclusive reliance on local features (insensitive to spatial relations, found for example for the Animals dataset and the scrambling-insensitive classes in ImageNet) to a stronger reliance on spatial relations (for example for the Sketchy dataset and the scrambling-sensitive classes in ImageNet). (3) We furthermore show to what extent spatial relations among features are used by CNNs to perform object recognition tasks. In particular, we provide evidence that the spatial arrangement of features is used only to construct features up to an intermediate level of granularity. That is, we did not find evidence of spatial integration in CNNs that allows them to capture the global shape of the objects in the datasets tested.

One possible explanation for a bias towards local features is the locality of the convolution operation (Baker et al., 2020). However, our finding that CNNs learned features of intermediate granularity for classification agrees with another possible explanation, namely that a bias to local features and not to global shape is a consequence of the optimization process for object classification (Malhotra, Dujmović, Hummel, & Bowers, 2022). Specifically, from an information-theoretic perspective, features of intermediate granularity are the most informative for image classification tasks (Ullman, Vidal-Naquet, & Sali, 2002). The idea is that, on the one hand, very complex features could be highly diagnostic because their presence gives high confidence about the class identity. However, on the other hand, these complex features may not be sufficiently sensitive (i.e. they do not exist in each exemplar) to be generalizable across exemplars of an object class. By contrast, very simple features would generalize better, but in addition would also lead to more false positives (i.e. lower specificity). Thus, features of intermediate complexity can provide an optimal trade-off between sensitivity and specificity (Ullman et al., 2002). Interestingly, it has been shown that randomly initialized CNNs display an increase in the representational structure similarity from early to late layers between different levels of abstraction of visual stimuli (photos, drawings, and sketches) (Singer, Seeliger, Kietzmann, & Hebart, 2022). However, after training CNNs on ImageNet, they showed a drop in the representational structure similarity in later layers after peaking in the intermediate layers which consequently led to lower classification performance on drawings and sketches. These results show that the optimization process and not the CNN architecture steers the representations to be biased to more local features that are optimal for solving its objective function. Along the same lines, prepending regular CNNs with a fixed non-trainable bank of Gabor filters led to better out-of-distribution generalization to line drawings, silhouettes, robustness to noise corruptions (Evans, Malhotra, & Bowers, 2022) and adversarial attacks (Dapello, Marques, Schrimpf, Geiger, Cox, & DiCarlo, 2020). These findings further suggest that similar to the sketchy dataset, limiting the amount of surface information through the Gabor filters led the CNNs to depend on more coarse-grained features that were more robust to pixel corruptions and more generalizable to different visual domains.

The bias towards local features can also be related to the idea of simplicity bias of neural networks, which states that neural networks preferentially extract the simplest features needed to solve a given task (Malhotra, Evans, & Bowers, 2020; Shah, Tamuly, Raghunathan, Jain, & Netrapalli, 2020). Consistent with this explanation, our MIRC analysis showed that models with small ERFs that by design are only capable of extracting simpler fine-grained features require larger patches of images for correct object recognition (because they have lower specificity). In contrast, models with larger ERFs that are capable of extracting more coarse-grained and more specific features were able to assign objects to their corresponding correct classes based on smaller image patches. Therefore, our results suggest that optimization for object recognition is unlikely to yield bias to the global shape of objects, even if the models have the capacity to learn it. A similar principle may hold for human vision, as it has been shown that in humans shape bias can be task- and context-dependent (Cimpian & Markman, 2005; Diesendruck & Bloom, 2003; Yoshida & Smith, 2003).

Our results have major implications for the ongoing discussion concerning shape and texture representation in CNNs, and

whether certain biases exist. There is little consensus about the extent to which CNNs are texture- or shape-biased. Some studies have suggested that CNNs are shape-biased (Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022), whereas others have suggested that CNNs are strongly texture-biased (Baker & Elder, 2022; Baker et al., 2018, 2020; Geirhos et al., 2019). Here, instead of using a shape-texture dichotomy to understand the nature of CNN representations, we have used the dichotomy of local vs. global features. We argue that this dichotomy is useful for two reasons: (1) It can be quantified without specific interpretations of what constitutes texture or shape, as we showed with our feature-scrambling approach. In fact, our approach does not test specific assumptions about the nature of the representations because we do not perform specific image manipulations to provide evidence for either texture or shape bias. Rather, we manipulate the network architecture and the spatial arrangement of the representations to determine the locality of the features. (2) It is flexible in that it allows local features to be both shape-like or texture-like. This means that the shape-texture dichotomy only maps partially to the global-local dichotomy. For example, this dichotomy is able to account for the existence of highly diagnostic shape features of fine granularity that are highly specific and sensitive (e.g., the nose of a dog). Indeed, when ranking ImageNet classes according to their scrambling sensitivity, it is not always obvious that the scrambling-sensitive classes would map to shape classes as would be intuitively expected. A possible explanation for previous inconsistent findings with respect to shape and texture is that the respective studies made very specific manipulations that did not generalize beyond these examples. For example, the texture bias observed in CNNs trained on ImageNet when tested on shape-texture cue conflict stimuli (Geirhos et al., 2019) was significantly reduced when the background of the images was removed (Tartaglini et al., 2022). Our findings suggest an explanation for these observations, in that the fine-grained (texture) features are less reliable than the more coarse-grained (shape) features, and therefore need to cover a large portion of the image to be diagnostic. Removing them from the background reduced their predictive power and led CNNs to be more shape-biased (Tartaglini et al., 2022) (on this specific test set). Another example is that many studies used silhouette stimuli to test shape bias in CNNs (Baker & Elder, 2022; Geirhos et al., 2019; Kubilius et al., 2016) and reached different conclusions. However, they used different datasets containing different classes. According to our results, this is expected since CNNs employ different classification strategies per object class and consequently will lead to variable classification performances on silhouette stimuli if the classes are different.

Given that CNN models are currently used as models of brain activity, specifically for the ventral stream of the visual system, which is believed to be responsible for object recognition (Cadieu et al., 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Yamins & DiCarlo, 2016; Yamins et al., 2014), it is important to understand the representations they develop and how they deviate from the brain. Although humans rely mostly on complex shape cues for object recognition (Landau et al., 1988), recent evidence has shown that the categorical organization of the entire ventral stream can be explained by mid-level features that do not include intact objects and do not convey any semantic information (Ayzenberg & Behrmann, 2022; Henderson, Tarr, & Wehbe, 2023; Jagadeesh & Gardner, 2022; Long, Yu, & Konkle, 2018). Moreover, albeit it is likely that humans rely on more than one mechanism to object recognition (Peissig & Tarr, 2007; Smith, 2009), some of these mechanisms might only depend on patchy diagnostic local features (Ullman, Sali, & Vidal-Naquet, 2001) especially given the fact that humans are capable of recognizing familiar objects from local image patches (Ullman et al., 2016) and these image patches evoke responses in higher-order category-selective visual areas (Holzinger, Ullman, Harari, Behrmann, & Avidan, 2019). Furthermore, it has been reported that human children's ability to recognize objects based on their global shape begins to develop only at 18–24 months of age (Pereira & Smith, 2009; Yee, Jones, & Smith, 2012). Before that, they are capable of recognizing objects based solely on their local features. In general, it has been shown that categorization of objects in humans relies on combinations of different perceptual and high-level semantic mental object representations constructed to model human similarity judgments (Hebart, Zheng, Pereira, & Baker, 2020). These results bear a resemblance to our findings of the heterogeneity of CNNs classification strategies across different datasets and different classes in ImageNet. The heterogeneity of CNN classification strategies across datasets also agrees with the observations in the literature that CNNs trained for object recognition rely on higher and wider distributions of spatial frequencies than CNNs trained on face recognition and consequently exhibited less robustness to blurring (Jang & Tong, 2021) and it is believed that humans recognize faces holistically as a whole in contrast to objects that can be recognized as a set of independent features (Grand, Mondloch, Maurer, & Brent, 2004; Tanaka & Simonyi, 2016). Our results, therefore, provide additional evidence for the hypothesis that features of intermediate granularity which are optimal for object recognition (Ullman et al., 2001, 2002) could be shared between CNNs and the ventral stream of the visual cortex (Henderson et al., 2023; Jagadeesh & Gardner, 2022; Long et al., 2018).

In summary, we showed here that although CNNs do not exploit global shape representations to perform object recognition, they can learn to utilize distributed feature constellations if this is required for solving the object classification task at hand. Looking ahead, we hypothesize that developing new tasks and objective functions to train CNNs instead of object recognition might lead to biases more aligned with humans. Reinforcement learning (RL) is a candidate objective function because it has been suggested that manual exploration may be a key factor in the development of shape bias in children (Pereira, James, Jones, & Smith, 2010; Soska & Johnson, 2008) and it has been shown that action planning using RL leads to divergent representation than supervised and unsupervised learning (Lindsay, Merel, Mrsic-Flogel, & Sahani, 2021). Moreover, neural agents that are trained to communicate efficiently i.e. be optimal on the trade-off between informativity and complexity of the messages used were shown to exhibit shape bias (Portelance, Frank, Jurafsky, Sordoni, & Laroche, 2021). Future investigations of such novel objective functions can not only lead to more effective biases and representations in such networks but also shed more light on how the observed human biases emerge.

5. Conclusions

We provide evidence that CNNs have the capacity to learn the spatial relations between features for object recognition. Specifically, the spatial arrangement of features is exploited by CNNs to build more coarse-grained features that are more reliable for object classification. Notably, the capacity of CNNs to learn the spatial arrangement of features varies according to the dataset and according to the class within the same dataset. We noticed, however, that CNNs employ the spatial configuration of features to build more coarse-grained features only up to an intermediate degree of granularity and do not exploit the global shape of objects. The reason for this is that features of intermediate granularity are more likely to be optimal in the trade-off between sensitivity and specificity i.e. generalizable and yet reliable.

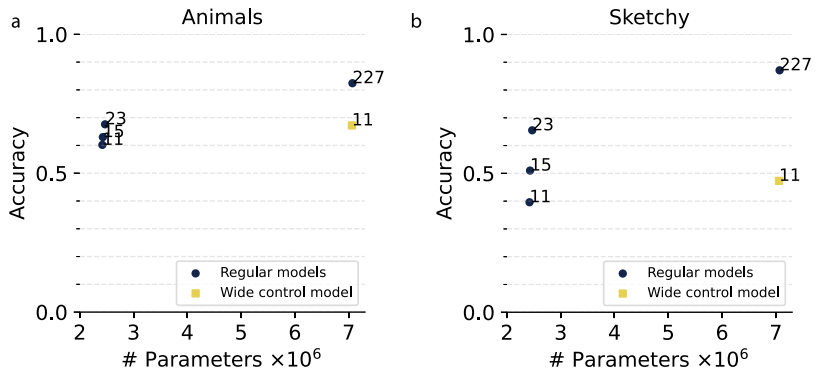

Fig. A.1. A control experiment in which we trained a wide model that has a small ERF (11 pixels), while matching the number of parameters of the model with the largest ERF (227 pixels). The numbers shown in the figure are the ERF of the corresponding models in pixels.

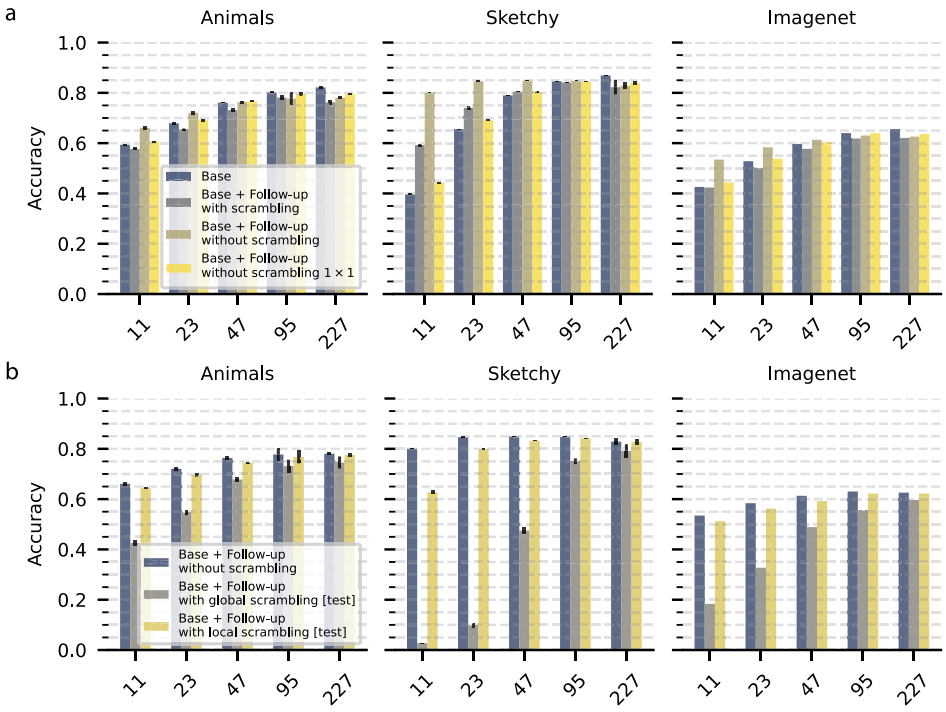

Fig. A.2. (a): Classification accuracy for CNN models of different ERFs under different training conditions of the feature-scrambling approach (Fig. 1d). (b): Classification accuracy of the base models with spatial aggregation without scrambling under different testing conditions (global and local scrambling).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this research is public

Acknowledgments

This project was supported by ERC Starting Grant to M.V. (SPATEMP) and a BMF Grant to M.V. (Bundesministerium fuer Bildung und Forschung, Computational Life Sciences, project BINDA, 031L0167).

Appendix

Controlling for the number of parameters of the models. We performed a control analysis to verify that the performance differences observed in our study among CNNs of different ERFs can be attributed indeed to their ERFs and not the number of model parameters. We trained a wider model of small ERF (11×11 pixels) but *with* matched the number of parameters to the model with the largest ERF (227×227 pixels). For both the Animals and Sketchy datasets, we observed a slight increase in the classification performance of the models by increasing the number of parameters. However, a small ERF model with a large number of parameters did not reach the performance of the model with the largest ERF, indicating the importance of the ERF to the models' performance. Furthermore, for the Sketchy dataset, the performance of the wider model with ERF $= 11 \times 11$ did not even reach the performance of the regular model with ERF $= 15 \times 15$ pixels. This is in line with our other results showing the reliance of the performance of CNNs on their ERF size,

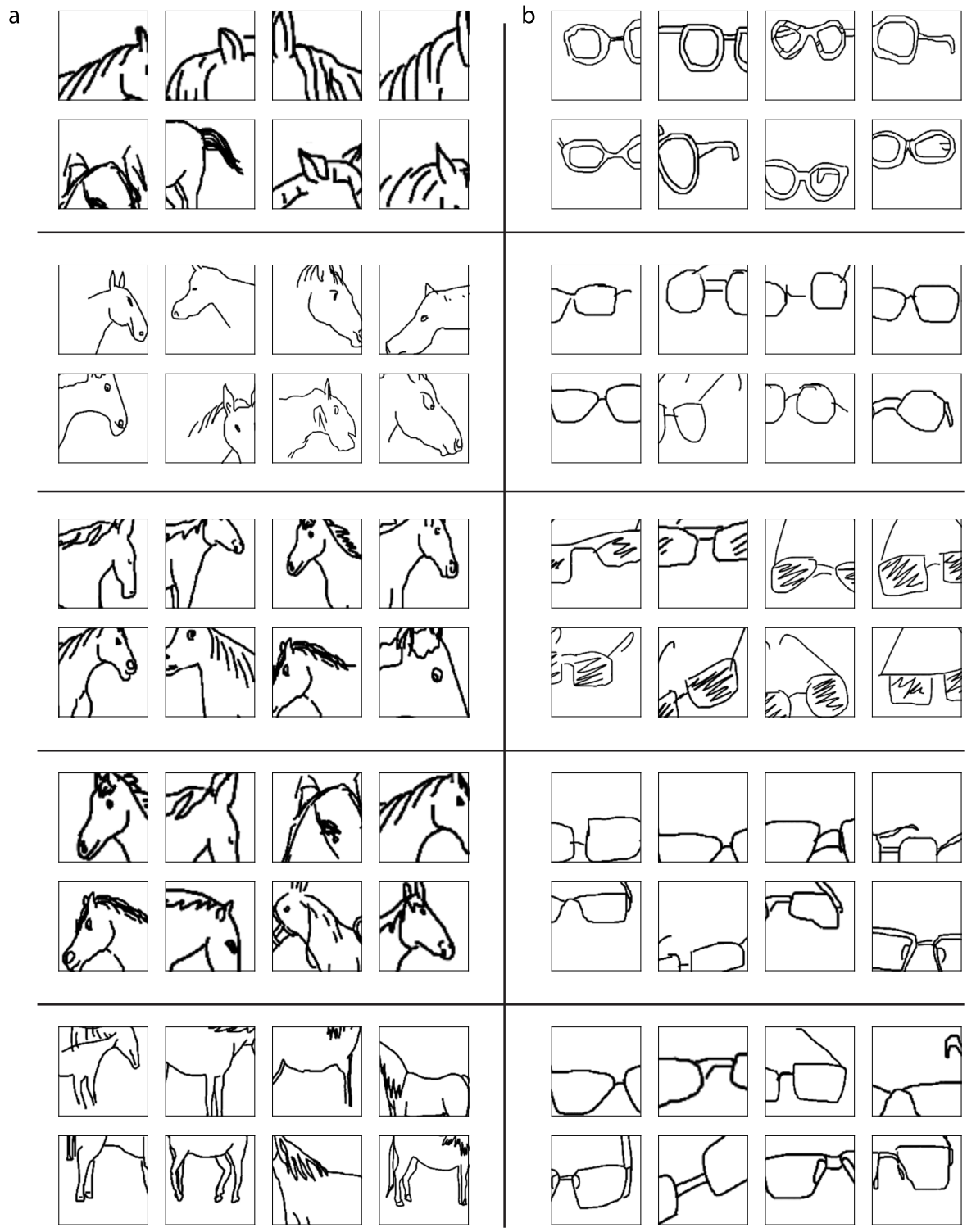

Fig. A.3. Clustering of all the MIRCs of the horse (a) and eyeglasses (b) classes (Sketchy dataset) in the representational space of the model ERF227 Each panel shows the eight closest MIRCs, generated from unique test images, in the representational space to the center of one cluster.

especially for the Sketchy dataset. Note that in the manuscript, we included several additional controls, e.g. scrambling during training, a 1×1 follow-up network, and local scrambling, which further show the importance of ERF size.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org, URL https://www.tensorflow.org/.

Ayzenberg, V., & Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*, http://dx.doi.org/10.1016/j.tics.2022.09.019.

Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience*, *25*(9), Article 104913. http://dx.doi.org/10.1016/j.isci.2022.104913.

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295. http://dx.doi.org/10.1037/xge0000409.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), 1–43. http://dx.doi.org/10.1371/journal.pcbi.1006613.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, *172*(October 2019), 46–61. http://dx.doi.org/10.1016/j.visres.2020.04.003.

Barenholtz, E., & Tarr, M. J. (2006). Reconsidering the role of structure in vision. *Psychology of Learning and Motivation*, *47*, 157–180. http://dx.doi.org/10.1016/S0079-7421(06)47005-5.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. http://dx.doi.org/10.1037/0033-295X.94.2.115.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64. http://dx.doi.org/10.1016/0010-0285(88)90024-2.

Brendel, W., & Bethge, M. (2019). Approximating NNs with Bag-of-Local-Features models works surprisingly well on ImageNet. In *7th International conference on learning representations, ICLR 2019* (pp. 1–15). arXiv:1904.00760.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), 1–35. http://dx.doi.org/10.1371/journal.pcbi.1003963.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. http://dx.doi.org/10.1038/srep27755.

Cimpian, A., & Markman, E. M. (2005). The absence of a shape bias in children's word learning. *Developmental Psychology*, *41*(6), 1003. http://dx.doi.org/10.1037/0012-1649.41.6.1003.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *33, Advances in Neural Information Processing Systems* (pp. 13073–13087). Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2020/file/98b17f068d5d9b7668e19fb8ae470841-Paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee, http://dx.doi.org/10.1109/CVPR.2009.5206848.

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168–178. http://dx.doi.org/10.1111/1467-8624.00528.

Edelman, S. (1993). Representing three-dimensional objects by sets of activities of receptive fields. *Biological Cybernetics*, *70*(1), 37–45. http://dx.doi.org/10.1007/BF00202564.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2022). Biological convolutions improve DNN robustness to noise and generalisation. *Neural Networks*, *148*, 96–110. http://dx.doi.org/10.1016/j.neunet.2021.12.005.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. http://dx.doi.org/10.1109/CVPR.2016.265.

Geirhos, R., Michaelis, C., Wichmann, F. A., Rubisch, P., Bethge, M., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International conference on learning representations, ICLR 2019* (pp. 1–22). arXiv:1811.12231.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., et al. (2021). Partial success in closing the gap between human and machine vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), *34, Advances in Neural Information Processing Systems* (pp. 23885–23899). https://proceedings.neurips.cc/paper_files/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *31*, In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Grand, R. L., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science*, *15*(11), 762–768. http://dx.doi.org/10.1111/j.0956-7976.2004.00753.x.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). A sequence of object-processing stages revealed by fmri in the human occipital lobe. *Human brain mapping*, *6*(4), 316–328. http://dx.doi.org/10.1002/(SICI)1097-0193(1998)6:4%3C316::AID-HBM9%3E3.0.CO;2-6.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition 2016-Decem* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. http://dx.doi.org/10.1038/s41562-020-00951-3.

Henderson, M. M., Tarr, M. J., & Wehbe, L. (2023). A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, *43*(22), 4144–4161. http://dx.doi.org/10.1523/JNEUROSCI.1822-22.2023.

Holzinger, Y., Ullman, S., Harari, D., Behrmann, M., & Avidan, G. (2019). Minimal recognizable configurations elicit category-selective responses in higher order visual cortex. *Journal of Cognitive Neuroscience*, *31*(9), 1354–1367. http://dx.doi.org/10.1162/jocn_a_01420.

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*(17), Article e2115302119. http://dx.doi.org/10.1073/pnas.2115302119.

Jang, H., & Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of Vision*, *21*(12), 6. http://dx.doi.org/10.1167/jov.21.12.6.

Jo, J., & Bengio, Y. (2017). Measuring the tendency of CNNs to learn surface statistical regularities. ArXiv abs/1711.11561.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *25*, In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. In M. Bethge (Ed.), *PLoS Computational Biology*, *12*(4), Article e1004896. http://dx.doi.org/10.1371/journal.pcbi.1004896.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. http://dx.doi.org/10.1016/0885-2014(88)90014-7.

Le, H., & Borji, A. (2017). What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks? arXiv preprint arXiv:1705.07049.

LeCun, Y., Yoshua, B., & Geoffrey, H. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Lindsay, G. W., Merel, J., Mrsic-Flogel, T., & Sahani, M. (2021). Divergent representations of ethological visual inputs emerge from supervised, unsupervised, and reinforcement learning. arXiv preprint arXiv:2112.02027.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(38), E9015–E9024. http://dx.doi.org/10.1073/pnas.1719616115.

Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. S. (2022). Human shape representations are not an emergent property of learning to classify objects. preprint http://dx.doi.org/10.1101/2021.12.14.472546.

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68. http://dx.doi.org/10.1016/j.visres.2020.04.013.

Margalit, E., Biederman, I., Tjan, B. S., & Shah, M. P. (2017). What is actually affected by the scrambling of objects when localizing the lateral occipital complex?. *Journal of Cognitive Neuroscience*, *29*(9), 1595–1604. http://dx.doi.org/10.1162/jocn_a_01144.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), http://dx.doi.org/10.1371/journal.pcbi.1003553.

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 Years ago? *Annual Review of Psychology*, *58*, 75–96. http://dx.doi.org/10.1146/annurev.psych.58.102904.190114.

Pereira, A. F., James, K. H., Jones, S. S., & Smith, L. B. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, *10*(11), 22. http://dx.doi.org/10.1167/10.11.22.

Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental Science*, *12*(1), 67–80. http://dx.doi.org/10.1111/j.1467-7687.2008.00747.x.

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The emergence of the shape bias results from communicative efficiency. (pp. 607–623). https://aclanthology.org/2021.conll-1.48.pdf,

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2002). The effect of image scrambling on visual cortical BOLD activity in the anesthetized monkey. *NeuroImage*, *16*(3 I), 607–616. http://dx.doi.org/10.1006/nimg.2002.1086.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience*, *38*(33), 7255–7269. http://dx.doi.org/10.1523/JNEUROSCI.0388-18.2018.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th international conference on machine learning - Volume 70, ICML '17* (pp. 2940–2949). JMLR.org, arXiv:1706.08606.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, *35*(4), http://dx.doi.org/10.1145/2897824.2925954.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, *33*, 9573–9585, https://proceedings.neurips.cc/paper/2020/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf.

Singer, J. J., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2022). From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of Vision*, *22*(2), 4. http://dx.doi.org/10.1167/jov.22.2.4.

Smith, L. B. (2009). From fragments to geometric shape: Changes in visual object recognition between 18 and 24 months. *Current Directions in Psychological Science*, *18*(5), 290–294, https://www.jstor.org/stable/20696051.

Soska, K. C., & Johnson, S. P. (2008). Development of three-dimensional object completion in infancy. *Child Development*, *79*(5), 1230–1236. http://dx.doi.org/10.1111/j.1467-8624.2008.01185.x.

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: A review of the literature. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1876–1889. http://dx.doi.org/10.1080/17470218.2016.1146780.

Tartaglini, A. R., Vong, W. K., & Lake, B. M. (2022). A developmentally-inspired examination of shape versus texture bias in machines. *Proceedings of the 44th annual conference of the cognitive science society*, arXiv:2202.08340.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2744–2749. http://dx.doi.org/10.1073/pnas.1513198113.

Ullman, S., Sali, E., & Vidal-Naquet, M. (2001). A fragment-based approach to object representation and classification. In *Visual form 2001* (pp. 85–100). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/3-540-45129-3_7.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687. http://dx.doi.org/10.1038/nn870.

Vogels, R. (1999). Effect of image scrambling on inferior temporal cortical responses. *NeuroReport*, *10*(9), 1811–1816. http://dx.doi.org/10.1097/00001756-199906230-00002.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194. http://dx.doi.org/10.1016/S0301-0082(96)00054-8.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2251–2265. http://dx.doi.org/10.1109/TPAMI.2018.2857768.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. http://dx.doi.org/10.1038/nn.4244.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. http://dx.doi.org/10.1073/pnas.1403112111.

Yee, M., Jones, S. S., & Smith, L. B. (2012). Changes in visual object recognition precede the shape bias in early noun learning. *Frontiers in Psychology*, *3*(DEC), 1–13. http://dx.doi.org/10.3389/fpsyg.2012.00533.

Yoshida, H., & Smith, L. B. (2003). Shifting ontological boundaries: how Japanese-and english-speaking children generalize names for animals and artifacts. *Developmental Science*, *6*(1), 1–17. http://dx.doi.org/10.1111/1467-7687.00247_1.